Modeling Data from Toronto

Dhanraj Patel

2021-08-02

Contents

Frequentist Model

Overview

The city of Toronto Prides itself on having one of the best public transportation systems in all of the world. While riding on one of the cities numerous TTC buses one is able to reach every corner of the City and while the city strives to keep their system running on a timely bases, there are situations which cause a delay. Using the data provided by the city of Toronto we will be modeling and estimating information about the delays that occur. Particularly, we will be estimating the rate of delays that occur in each day of the week per 6 hour intervals.

Data Collection Process

Before delving deeper into the data it is important to understand the process of how the data was collected. Whenever a TTC bus delay occurs, the TTC (Toronto Transit Commission - the public agency in charge of transportation in Toronto) records the event in detail. The TTC makes sure to record the time and day of the delay, the route of the delayed bus, location of delay, the amount of time the bus was delayed for and the cause of the delay. Due to the fact that the TTC is a public agency all the information they record is then passed on to the city of Toronto who then publish it for the public to view. The data is available on https://open.toronto.ca.

Cleaning of data

After retrieving the data set it was important to clean the data so that the results generated would be an accurate reflection of the data. In order to clean the data it was first important to eliminate all the duplicate entries. Within the data set there were instances were the same delay was recorded multiple times. For the purposes of this analysis, since we only want to include one recorded instance of all delays, all the duplicates in the data set were removed.

In addition, within the data set both of the columns of route and direction of the recorded delayed bus contained missing values. As for the direction column about 10 percent the values were missing and for route about 0.5 percent of the values were missing. Due to the fact that the values for route and direction were not used for the the models and estimations, the missing values for both columns were not removed as not not alter the accuracy of the data that would be caused by removing data.

Lastly, since the delay times were recorded down to the minute, I introduced a new variable called time_interval. With this, I split up the time into 4 sections so that the data can be modeled to find rate of delays that occur in 6 hour intervals.

Comparing rate of delays by days of the week

In statistics we are able to use models and estimate true population parameters that otherwise would be very difficult to find. From this, we are able to model the rate at which delays occur in the different days of the week. When discussing the rate which something occurs its important to select an appropriate time frame that the rate is measured in. For the purposes of this model we will be modeling the rate in a time frame of 6 hours, so that each day of the week can be sectioned into 4 intervals. Using this information the most appropriate to model to estimate the rate at which delays occur in each day of the week is the following:

The Model

$$X_{i,j} \sim Pois(\lambda_j)$$

In this model $X_{i,j}$ is a random variable denoting the number of delays in the day of the week i. In addition, $Pois(\lambda_j)$ is the estimate rate of the average number of delays in the day of the week i per 6 hour interval.

Explanation of methodologies

From our data we are able to record for each day of the week, the number of delays that occurred in a time interval of 6 hours ($X_{i,j}$). The intervals are broken down into time of day from:

- Morning: 00:00 to 06:00
- Afternoon: 06:01 to 12:00
- Evening: 12:01 to 18:00
- Night: 18:01 to 24:00

Using the data our goal is to estimate the average number of delays in the day of the week i per 6 hour interval. Since we are using our data to draw conclusions of an estimate by emphasizing the frequency or proportion of the data and our parameter of interest is an unknown numerical value, the methodology we are using to attain the estimate is of the frequentist statistics type.

As for our model it is important to note that since the rate is measuring a number of events (number of delays) in a time interval and the fact that each event is discrete (the measure of number of delays is discrete). Due to this the best distribution family to use for our model is the Poisson distribution.

Assumptions

As with many model, there are a number of assumptions to consider. Our main source of assumptions lies with the fact that we are using a Poisson distribution, in order to use our model we need to assume that:

- No more than one event can occur simultaneously
- Events occurring in different intervals are independent
- Expected number of events in each time interval is constant

Parameter of interest

For our model the parameter of interest that we are attempting to estimate is λ_j, the parameter represents rate of the average number of delays in the day of the week i per 6 hour interval. Since we are using the frequentist method our parameter of interest is an unknown numerical value that we will estimate using our data.

Aggregation of data

Table 1: Count of observed delays within 6 hour intervals

Mon_delays	Tue_delays	Wed_delays	Thu_delays	Fri_delays	Sat_delays	Sun_delays
64	71	71	66	65	24	11
71	71	71	80	72	34	60
4	10	13	8	15	6	7
31	42	42	33	33	33	21
52	74	81	67	38	23	35
48	77	69	70	61	39	32
11	10	16	12	5	10	9
28	34	32	35	27	22	20
76	62	72	89	52	47	39
55	64	78	74	80	69	55
12	9	8	11	8	12	8
37	34	26	30	42	32	28
73	53	56	64	73	43	35
65	51	48	62	63	55	57
11	12	6	11	8	16	11
30	26	33	24	27	22	33
NA	74	86	NA	NA	NA	NA
NA	67	78	NA	NA	NA	NA
NA	7	14	NA	NA	NA	NA
NA	23	34	NA	NA	NA	NA

From (Table_1) are able to see in the month of July 2021 the number of delays recorded in each 6 hour interval for every day of the week. It is important to note that there was one more Wednesday and Tuesday in the month compared to every other day, so for Tuesday and Wednesday there are 4 more rows of data for that extra day. All other days in the last 4 rows show NA. We are able to use this data to estimate the value of $Pois(\lambda_j)$ for every day of the week.

Results - Comparing rate of delays by days of the week

With the data all in (Table_1) we are able to $Pois(\lambda_j)$ using the maximum likelihood estimator. The data observed in June 2021 can be used to estimated by finding the value that maximizes the probability of the collected data. Using this method, the estimates are:

Table 2: Average number of delays in day within 6 hours

Day	Num_of_delays
Monay	42
Tuesday	44
Wednesday	47
Thursay	46
Friday	42
Saturday	30
Sunday	29

In (Table_2) we see the results of the average number of delays within a six hour interval for every day of the week. With the value of the number of delays corresponding to the estimate $\hat{\boldsymbol{\lambda}}_{\boldsymbol{j}}$ for each day of the week.

Interpreting of the results

From (Table_2) we are able to see the results of our estimates. Looking at the table are able to see that Wednesday has the highest count of the average number of delays in a six hour interval while Sunday has the lowest value. In addition, we can see that the averages for every weekday are close together, and the averages of Saturday and Sunday are also close together. Comparing these two, the averages of the weekdays are noticeably higher than the weekends. I believe this can be explained with the fact that many people use TTC buses in their commute to work and thus with the TTC most busy on the weekdays, it increases the average number of delays. In the weekends, when the traffic on the TTC buses is lowest, there is in turn a much lower average for the number of delays. From our results we are able to conclude that the average number of delays in a six hour interval, is higher in the weekdays (highest on Wednesday) and lower on the weekends (lowest in Sunday) with a large factor in these results being population traffic and how busy the TTC buses are.

In addition, from what we can see, the results do seem reasonable. We know that they are reasonable by comparing the estimation for each day of the week, with the data collected in (Table_1). Comparing these two, we see that in (Table_1) the average of the number of delays in the data was highest in Wednesday and lowest in Sunday. This correlates to our estimates in (Table_2) where Wednesday had the highest value and Sunday had the lowest value. A similar trend can be noticed where in both tables the averages for the weekdays were noticeably higher than the weekends. The values of the estimates are reasonably set within expectations given by the data in (Table_1). Lastly, we know that the estimates are intuitively reasonable, due to the fact that on weekdays. Due to commuting the foot traffic traffic on the TTC buses is higher compared to the weekends, leading to increased delays.

Linear Regression Model

Overview

Whenever an assault related case is reported to the TPS (Toronto Police service) the crime is investigated and recorded in great detail. The recorded assault case is then further sub divided into the categories of: breaking and entering, automotive theft and robbery so that more detail of what types of assaults are occurring in the city can be investigated. In addition, the TPS also records which neighborhood the assault cases occur in to gain a better understanding of which regions in Toronto are experiencing the most assault crimes. The TPS gathers all the data and release the information annually and include all the data collected over the last 7 years. This data set contains the information collected from 2014 to 2020. Using this data, we can use linear regression modeling to see analyze how the number of assault related crimes have been changing year by year.

Data Collection Process

Before analyzing how the number of assault related crimes change from 2014 to 2020, it's important to understand how the data was collect. Whenever a assault related crime is reported, the TPS record the event in detail. They record the date of the crime, type of assault crime (breaking and entering, robbery or theft of automotive vehicle), and the neighborhood that the crime occurred in. The TPS then in in give the data to the Toronto municipal government which then in turn makes the data available to the public, the data can be found on https://open.toronto.ca.

Cleaning of data

After attaining the data, the data set was not changed or altered in any way to clean the data. The data set contained no missing values or duplicates that would need to be potentially removed.

Comparing number of assault cases from 2014-2020

In statistics we are able to using linear regression modeling to analyze how the number of assault related cases have been changing from the year 2014 up to 2020. To accomplish this we will be using the model:

The Model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Where x_i is the ith year, and Y_i is the total number of reported assault related cases in the ith year. Will be estimating β_1 which will indicate the trend of the number of assault cases from 2014 to 2020. $\epsilon_i \sim N(0, \sigma^2)$ describes the component random component of the linear relationship between x_i and Y_i. the value of β_0 can be interpreted as the y-intercept of the model, the value when the x_i is equal to 0.

Explanation of methodologies

From our data set we are able to extract the the total amount of assault related cases reported in each year from 2014 till 2020. From this information we are able to use the linear regression model to estimate β_1 which can be seen as the "line of best fit" of our data. With β_1, we will be able to analysis the trends occurring from 2014 till 2020 and gain an understanding of whether or not the number of total assault cases are increasing or decreasing and to what degree.

Assumptions

As with many model, there are a number of assumptions to consider. Our main source of assumptions lies with the fact that we are using linear regression model, in order to use our model we need to assume:

- Linearity: The relationship between x_i and the mean of Y_i is linear.
- Homoscedasticity: The variance of residual is the same for any value of x_i.
- Independence: Each one of the observations are independent of each other.
- Independence: Normality: For any fixed value of x_i, Y_i is normally distributed.

Parameter of interest

In our model we will be estimating β_1, this estimate can be seen as the "line of best fit" of our data which contain the of number of assault related cases from 2014 till 2020. With this value we will be able to understand the trend of the total number of assault cases from 2014 till 2020.

Aggregation of data

Table 3: Total number of assault cases per year

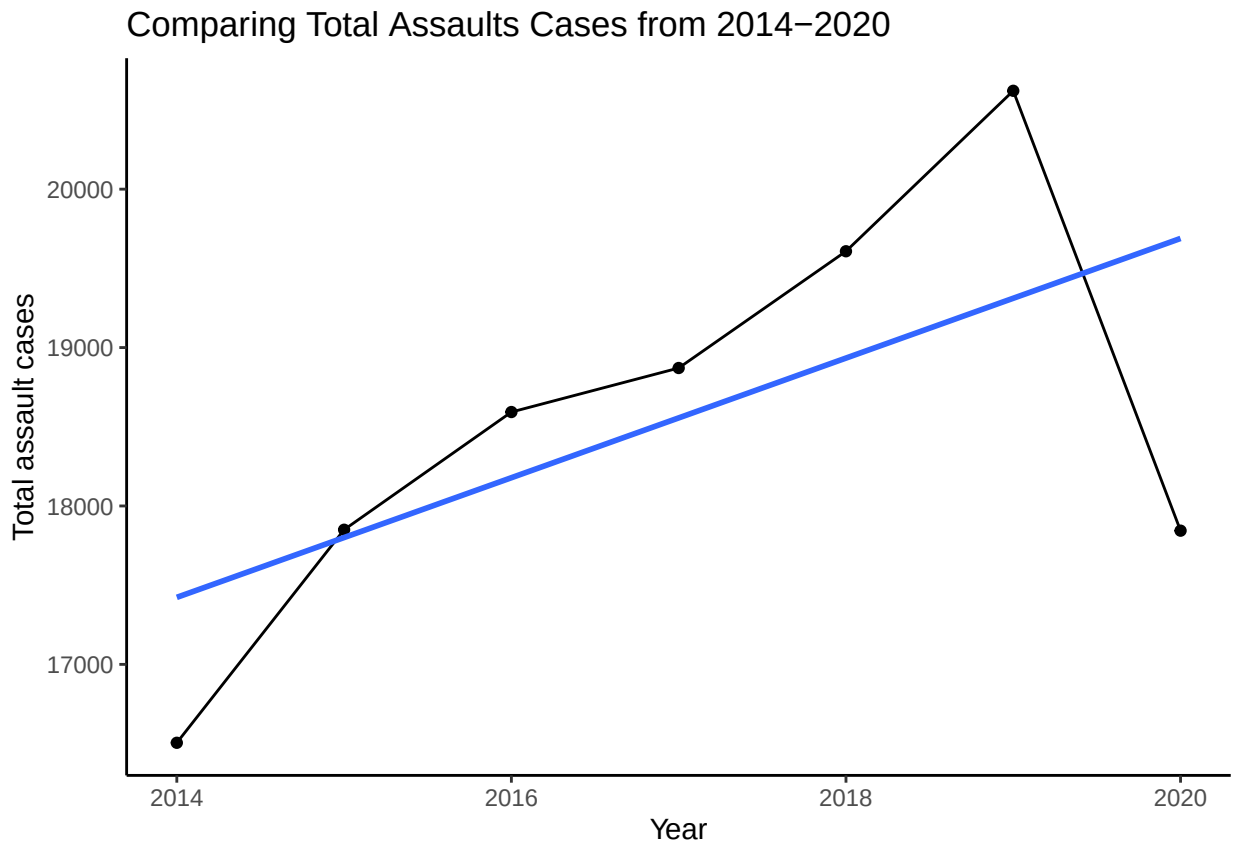
Year	Total_assault_cases
2014	16505
2015	17850
2016	18593
2017	18871
2018	19608
2019	20621
2020	17844

From (Table_3) we are able to to see the total amount of assault related cases that occurred in Toronto from 2014 to 2020. We can then graph the data in a scatter-plot and then use our linear regression model to estimate a line of best fit which will help us understand the trend of assault cases through the years. From looking at the data we are able to see that from all the year starting with 2014 and up until the year 2019 the number of total assault cases have increased. The year 2020 is the only year we see a decrease in the total number of assault cases

Results - Comparing number of assault cases from 2014-2020

With the data all in (Table_3) we are able to use linear regression to estimate a "line of best fit" to understand the trends in the total number of assault related cases in Toronto, using this value, the estimate β_1 is 377.6. As such, the estimated increase in assault cases from 2014 to 2020 year over year is 377.6 cases. To gain a better understanding of the data we can visualize the results with a graph.

(Graph_1)



From (Graph_1) we are able to see 2 lines. The black line represents the data collected of the total number of assault cases recorded from the years 2014 to 2020. The blue line represents the line of best fit with the slope of final estimate of β_1, which was 377.6. From this graph we are able to see that from the years 2014 to 2020, the assault cases in Toronto increased an average of 377.6 cases per year.

Interpreting of the results

From (Graph_1) we are able to see that on average the number of assault cases increased by 377.6 cases per year. We are able to to interpret that Toronto is experiencing an increasing number of assault cases every year on average. From the collected data (black line) we are able to that from 2014 up until 2019, every year has had an increasing number of assault cases. This increase is caused by a multitude of different reasons, such as an increasing population in the City. Since every year, the number of Toronto residents increases, it also follows that the number of assault cases also rise along with it. In the data however, we see a significant drop in the number of assault cases from 2019 to 2020, this drop impacted β_1 greatly by decreasing its estimated value and in turn the slope of the line of best fit. This decrease also causes an underestimation of the average increase in assault cases from the years 2014 to 2019. The fall in cases in 2020 can be explained by the world-wide Coronavirus pandemic which resulted in the city of Toronto going into strict lockdowns. With many people isolating in their homes, this caused a significant decrease of assault cases.

In addition, from what we can see, the results do seem reasonable. We know that they are reasonable by comparing the estimation for each year, with the data we collected in (Table_3). From the table we can see that the assault cases increased in every year, with a noticeable decrease in 2020 compared to 2019. From this is reasonable that the estimation of β_1 would indicate an increasing trend in total assault cases. Lastly, we know that the estimates are intuitively reasonable, due to the fact that due to an increasing population and other socioeconomic factors we are seeing an increase in assault cases every year and that due to the Coronavirus pandemic, there would be a noticeable decrease in total assault cases in 2020.

Resources

This entire document and all its contents were made by using the programming language R and Rstudio. From pulling the data, to cleaning the data, manipulating the data and creating tables and graphs, was all done using R and Rstudio.